DOCUMENT RESUME

ED 175 920 TM 009 562

AUTHOR Brennan, Robert L.; Lockwood, Robert E.
TITLE A Comparison of Two Cutting Score Procedures Using Generalizability Theory. ACT Technical Bulletin No. 33.
INSTITUTION American Coll. Testing Program, Iowa City, Iowa. Research and Development Div.
PUB DATE Apr 79
NOTE 63p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Achievement Tests; Comparative Statistics; *Cutting Scores; Guessing (Tests); *Item Analysis; *Mathematical Formulas; Multiple Choice Tests; Probability; Statistical Analysis; Test Construction; *Test Reliability
IDENTIFIERS *Angoff Method; Generalizability Theroy; Interrater Reliability; *Nedelsky Method

ABSTRACT

Procedures for determining cutting scores have been proposed by Angoff and by Nedelsky. Nedelsky's approach requires that a rater examine each distractor within a test item to determine the probability of a minimally competent examinee answering correctly; whereas Angoff uses a judgment based on the whole item, rather than each of its components. The reliability of these approaches depends upon the extent to which raters agree in their judgments. Generalizability theory was used to quantify the magnitude of error variance in each procedure; to compare data resulting from each procedure; and to examine the impact of rater disagreement on test reliability. Five subject experts rated the probability of answering correctly for a total of 126 four-option items in a health-related area. Both procedures were used by the same raters. Cutting score was assumed to be the observed mean (probability) over raters and items. In this sense, the expected variability of the observed mean was error variance attributable to the procedure used. Results indicated that both the cutting scores and their expected variance were considerably different for the two procedures, and suggested that differences between the procedures may be of greater consequence than their apparent similarities. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED175920

ACT TECHNICAL BULLETIN NO. 33

A Comparison of Two Cutting Score Procedures Using Generalizability Theory

by

Robert L. Brennan and Robert E. Lockwood

American College Testing Program

Iowa City, Iowa 52243

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert L. Brennan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Research and Development Division

The American College Testing Program

P. O. Box 168, Iowa City, Iowa 52243

April 1979

TM009 562

A Comparison of Two Cutting Score Procedures Using Generalizability Theory

by

Robert L. Brennan and Robert E. Lockwood

The American College Testing Program

A paper presented at the Annual Meeting
of the National Council on Measurement
in Education, San Francisco, April 1979.

Abstract

Nedelsky and Angoff have suggested procedures for establishing a cutting score based on raters' judgments about the likely performance of minimally competent examinees on each item in a test. In this paper, generalizability theory is used to characterize and quantify expected variance in cutting scores resulting from each procedure. Data for a 126-item test are used to illustrate this approach and to compare the two procedures. Finally, consideration is given to the impact of rater disagreement on some issues of measurement reliability or dependability. Results suggest that the differences between the Nedelsky and Angoff procedures may be of greater consequence than their apparent similarities.

A Comparison of Two Cutting Score Procedures Using Generalizability Theory

by

Robert L. Brennan and Robert E. Lockwood
American College Testing Program

TABLE OF CONTENTS

	Page
Introduction	1
Nedelsky and Angoff Procedures	1
Issues, Approach, and Data Sets	2
The r x i Design and the Angoff Procedure	5
Sample Statistics	6
Generalizability Theory	7
Generalizability Results for Angoff Procedures	9
The Nedelsky Procedure	12
Probabilities of Correct Responses	12
Eliminated Distractors	13
Eliminating Correct Alternative	16
A Comparison of the Two Procedures	17
Differences in Sample Statistics for Raters	19
Correlations and Covariances Among Raters, Within Procedures	20
Rater Means	20
Rater Standard Deviations	21
Differences in Sample Statistics for Items	22
Operationalizing Conceptions of Minimum Competence	24
Cutting Scores Other Than $\bar{X}$	26
Reconciliation Process	27
Nedelsky's Cutting Score	28
Measurement Reliability or Dependability	30
Summary and Conclusions	35
Appendix	37
References	41
Footnotes	43
Tables	44
Figures	55

Introduction

Currently, there is considerable debate concerning the setting of passing standards when scores on tests are used to make decisions regarding minimal competency, proficiency, licensure, certification, and the award of credit (see, for example, NCME, 1978). Meskauskas (1976), Buck (1977), and Zieky and Livingston (1977), among others, have reviewed current procedures for establishing cutting scores. For the most part, these procedures can be grouped into two categories--procedures based on subjective judgments of subject matter specialists, and procedures that use examinee scores on the test itself and/or some criterion measure. The latter procedures are not discussed in this paper; rather, primary emphasis is placed upon studying two procedures suggested by Nedelsky (1954) and Angoff (1971) for establishing cutting scores based upon the judgments of subject matter experts.

Nedelsky and Angoff Procedures

Both of these procedures require judgments by raters concerning the performance of hypothetical minimally competent examinees on each item of a test. The approach described by Nedelsky (1954) requires that a rater examine each distractor within an item to determine the probability that a minimally competent examinee would answer that question correctly; whereas the approach described by Angoff (1971) makes use of a judgment based on the whole item rather than its individual components.

Using Nedelsky's procedure, raters are asked to identify, for each item, those distractors that a minimally competent examinee would eliminate as incorrect. The reciprocal of the number of remaining alternatives (including

the correct answer) serves as an estimate of the probability that a minimally competent examinee would get the item correct; and the mean of these item probabilities, over items and raters, is defined as the cutting score for the test (in terms of proportion of items correct). In Angoff's procedure, raters simply provide an estimate of the item probabilities without specifically identifying which distractors a minimally competent examinee would eliminate. Again, the mean of these item probabilities, over items and raters, is defined as the cutting score for the test. Notationally, throughout this paper, we use $\bar{X}$ to denote the cutting score, or more specifically, the mean cutting score that results from a particular study. For a particular rater, r, the mean of that rater's item probabilities will be denoted $\bar{X}_r$, which can be interpreted as the cutting score that would be assigned by that particular rater.

Issues, Approach, and Data Sets

The Nedelsky and Angoff procedures are appealing in many contexts because they are understandable to raters and test users; and these procedures force raters to give detailed consideration to the specific content of a test, rather than to its general characteristics. However, the validity and practical utility of these approaches, and similar approaches, may rest heavily upon the extent to which raters agree in their judgments. This issue has received very little attention in the context of establishing cutting scores, although Andrew and Hecht (1976) do address some aspects of this issue. The principal purposes of this paper are: (a) to identify a psychometric approach for characterizing and

quantifying the magnitude of error variances (in either cutting score procedure) attributable to disagreement evident in rater judgments; (b) to apply this approach to data resulting from the Nedelsky and Angoff procedures, and thereby compare the two procedures; and (c) to examine the impact of rater disagreement on some issues relating to the reliability or dependability of measurement.

The principal psychometric approach employed to address these issues is based upon generalizability theory, although some aspects of these issues are addressed in more traditional ways. Generalizability theory (see Cronbach, Gleser, Nanda, and Rajaratnam, 1972) is especially appropriate here because it allows us to differentiate among multiple sources of error in a systematic manner. In the body of this paper we introduce and explain concepts and equations from generalizability theory, as needed; but we do not usually prove results. Readers desiring more detail are referred to Cronbach et al. (1972) and/or Brennan (1977). It should be noted that there are many aspects of generalizability theory that do not concern us in this paper. For example, we never report a generalizability coefficient and in only one instance do we refer to a universe score variance, as this term is defined in generalizability theory. Indeed, the approach used here is essentially variance components analysis viewed from the perspective of generalizability theory.

We employ three data sets to illustrate our approach to issues of rater disagreement and to compare the Nedelsky and Angoff procedures. Data set 1 consists of the Angoff probabilities assigned by five raters to each of the 126 four-alternative items constituting a test in a health-related area. Data set 2 consists of the (inferred) Nedelsky probabilities assigned by the same

raters to the same items; and data set 3 consists of the eliminated distractors, for each item and rater, that form the basis for inferring the Nedelsky probabilities in data set 2. Each of the five raters is a practitioner or teacher in the appropriate field; and, their ratings were provided independently. At the conclusion of the study, raters were instructed to discuss each item and provide a consensus Angoff-type judgment. These reconciled judgments are examined, although not extensively, in a separate analysis.

Both the Nedelsky and Angoff procedures necessitate judgments about "minimum competence." In one section of this paper, we briefly consider some aspects of how the two procedures allow a rater to operationalize some conception of minimum competence. Otherwise, however, this paper is not intended to treat educational, philosophical, or psychological issues associated with defining minimum competence. Also, throughout this paper, except in one section, we restrict ourselves to consideration of $\overline{X}$ as a cutting score. Finally, we recognize that in realistic settings evaluations sometimes use more than one cutting score procedure, or a variant of the procedures discussed here. This paper is not intended to address such issues in any detail.

The r x i Design and the Angoff Procedure

For the Angoff procedure, the probability assigned by the rater r to item i can be represented as

$$X_{ri} = \lambda + \lambda_r^{\sim} + \lambda_i^{\sim} + \lambda_{ri}^{\sim}; \qquad (1)$$

where λ = grand mean for the population of raters and the universe of items,

$\lambda_r^{\sim}$ = effect for rater r,

$\lambda_i^{\sim}$ = effect for item i, and

$\lambda_{ri}^{\sim}$ = effect for the interaction of rater r and item i.

(Technically, since we have only one observation for each rater-item combination, the effect $\lambda_{ri}^{\sim}$ is completely confounded with any other sources of variation--sometimes called "random" or "experimental" error.) Here, unless otherwise noted, we will assume that the actual raters in the study can be considered a random sample from an essentially infinite population of raters; and, that the actual items can be considered a random sample from an essentially infinite universe of items. Under this assumption, and assuming independent effects that sum to zero, Equation 1 represents what is usually called a random effects model for the r x i design.

Given this model, for rater r, the average probability over the universe of items is

$$\lambda_r = \lambda + \lambda_r^{\sim};$$

whereas the average probability over the sample of n_i items is $\bar{X}_r$. Similarly, for item i, the average probability over the population of raters is

$$\lambda_i = \lambda + \lambda_i \nu;$$

and the average probability over the sample of n_r raters is $\bar{X}_i$.

Sample Statistics

In terms of sample statistics, Table 1 reports means, standard deviations and intercorrelations among raters, for data set 1 and the Angoff procedure. In Table 1 and subsequent tables, all results, except those within parentheses, are in terms of probabilities. Results within parentheses are in terms of number of items. For example, Table 1 reports that the mean probability over the $n_r = 5$ raters and $n_i = 126$ items in this study is $\bar{X} = 0.6632$. In effect, this average probability is the (mean) cutting score, in terms of proportion of items correct, arrived at using the Angoff procedure. In terms of number of items correct, the (mean) cutting score is $n_i \bar{X}$, or 83.56, as reported in Table 1. Also, Table 1 reports that the standard deviation of the rater mean probabilities is 0.0373. This is the standard deviation of the cutting scores for the five raters, in terms of proportions of items correct. The corresponding standard deviation in terms of number of items correct is 4.70.

Insert Table 1 about here

We will examine the results reported in Table 1 in somewhat more detail, later. Here, we simply note that Table 1 suggests that there is some degree of variability among rater means, as reflected by $\hat{\sigma}(\bar{X}_{\underline{r}})$; there is some degree of variability within each rater, as reflected by $\hat{\sigma}(\bar{X}_{\underline{r}\underline{i}})$; and there is some degree of variability in the rater intercorrelations. The sample statistics in Table 1, however, do not indicate clearly the variability in the mean cutting score, $\bar{X}$, which is a principal concern of this paper. In other words, we would like some estimate of the variance (or standard deviation) of $\bar{X}$ if the entire study were replicated with different samples of raters and/or items. To obtain such estimates we employ generalizability theory.

Generalizability Theory

Given the random effects model in Equation 1, Table 2 reports equations for estimating the variance components associated with each of the score effects in the model. For example, $\hat{\sigma}^2(\underline{r})$ is an unbiased estimate of the variance of $\lambda_{\underline{r}}$ (or $\lambda_{\underline{r}}\sim$) over the <u>population</u> of raters. (Recall that $\lambda_{\underline{r}}$ is the expected value, over the <u>universe</u> of items, of the probabilities assigned by rater $\underline{r}$.) Similarly, $\hat{\sigma}^2(\underline{i})$ is an unbiased estimate of the variance of $\lambda_{\underline{i}}$ (or $\lambda_{\underline{i}}\sim$) over the universe of items.

It is important that $\hat{\sigma}^2(\underline{r})$ be differentiated from $\hat{\sigma}^2(\bar{X}_{\underline{r}})$. The former is an estimate of the variance, over the population of raters, of the scores (or probabilities) $\lambda_{\underline{r}}$; while the latter is the variance, over the sample of raters, of the scores (or probabilities) $\bar{X}_{\underline{r}}$. In terms of the random effects variance components in Table 2,

$$\hat{\sigma}^2(\bar{X}_r) = \hat{\sigma}^2(r) + \hat{\sigma}^2(ri)/n_i \quad . \qquad (2)$$

In other words, the observed variance of rater means over the $n_i = 126$ items can be decomposed into two parts--one part that is uniquely associated with raters, and another part that is associated with the interaction of raters and items.

Insert Table 2 about here

Table 2 also reports three equations for estimating the expected variance of $\bar{X}$. Each of these equations is expressed solely in terms of random effects variance components and sample sizes. The sample sizes in these equations are identified with primes to distinguish them from the sample sizes that characterize the actual data available. We say that n represents a G study sample size and n' represents a D study sample size. In the body of this paper, unless otherwise noted, we will assume that the G study and D study sample sizes are equal. In an Appendix we provide a more detailed consideration of distinctions between G studies and D studies for the $r \times i$ design.

Equation 3 in Table 2 provides the expected value of the variance of the mean cutting score, $\bar{X}$, for generalizing over samples of n'_r raters and n'_i items. We can conceive of the possibility of determining $\bar{X}$ a "very large" number of times--each time using a different sample of n'_r raters and n'_i items. Equation 3 estimates the variance of the distribution of the "very large" number of means that would result from such replications.

It is in this sense that we say $\hat{\sigma}^2(\underline{\bar{X}})$ is the variance of the mean for generalizing over both samples of raters and samples of items.

Equation 4 in Table 2 is the expected variance of $\underline{\bar{X}}$ generalizing over samples of $\underline{n_i'}$ items, for a fixed set of $\underline{n_r'}$ raters. We dencte this variance $\hat{\sigma}^2(\underline{\bar{X}}|\underline{P}^*)$ to emphasize that raters are considered fixed. Again, we can conceive of the possibility of determining $\underline{\bar{X}}$ a "very large" number of times--each time using a different sample of $\underline{n_i'}$ items but the same $\underline{n_r'}$ raters. $\hat{\sigma}^2(\underline{\bar{X}}|\underline{R}^*)$ is an unbiased estimate of the variance of this distribution of means. Similarly, $\hat{\sigma}^2(\underline{\bar{X}}|\underline{I}^*)$ in Equation 5 is the expected variance of $\underline{\bar{X}}$ generalizing over samples of $\underline{n_r}$ raters, for a fixed set of $\underline{n_i}$ items.

In brief, $\hat{\sigma}^2(\underline{\bar{X}}|\underline{I}^*)$ is for generalizing over samples of raters, $\hat{\sigma}^2(\underline{\bar{X}}|\underline{R}^*)$ is for generalizing over samples of items, and $\hat{\sigma}^2(\underline{X})$ is for generalizing over samples of both raters and items. These then are three different estimates of error variance in the mean cutting score. Which of these estimates is appropriate can be determined only in the context of a specific study; i.e., it is the decision-maker who must determine whether it is appropriate to generalize over samples of raters, items, or both. It is evident from Equations 3 to 5, however, that $\hat{\sigma}^2(\underline{\bar{X}})$ must be at least as large as $\hat{\sigma}^2(\underline{\bar{X}}|\underline{R}^*)$ and $\hat{\sigma}^2(\underline{\bar{X}}|\underline{I}^*)$. This follows from the fact that $\hat{\sigma}^2(\underline{\bar{X}}|\underline{R}^*)$ does not involve variability due to raters, $\hat{\sigma}^2(\underline{r})$, and $\hat{\sigma}^2(\underline{\bar{X}}|\underline{I}^*)$ does not involve variability due to items, $\hat{\sigma}^2(\underline{i})$.

Generalizability Results for Angoff Procedure

For data set 1 and the Angoff procedure, Table 3 reports the usual ANOVA results, estimated random effects variance components, and estimates of mean cutting score variability. As reflected by the above discussion, it is usual

in generalizability theory to report results in terms of variances; however, in Table 3 we also report the three estimates of mean score variability in terms of standard deviations to facilitate interpretation. We note, for example, that in terms of proportion of items, the standard deviation of $\overline{X}$, for generalizing over raters and items, is 0.0182; and in terms of number of items, it is 2.29. Furthermore, $\hat{\sigma}(\overline{X})$ and $\hat{\sigma}(\overline{X}|I^*)$ have approximately the same magnitude; and both of them are almost twice as large as $\hat{\sigma}(\overline{X}|R^*)$. Clearly, for these data, the decision concerning whether or not to generalize over raters is an important determiner of the magnitude of the standard deviation of $\overline{X}$.

Insert Table 3 about here

The results reported in Table 3 are based on the assumption that the G and D study sample sizes are the same; i.e., $n_r = n'_r = 5$ and $n_i = n'_i = 12$. The equations in Table 2 can be used, however, to determine the expected variability of $\overline{X}$ for different numbers of raters and/or items. For example, the reader can verify that, if the number of raters were doubled to $n'_r = 10$ and the number of items remained unchanged, then the values of $\hat{\sigma}(\overline{X})$, $\hat{\sigma}(\overline{X}|R^*)$, and $\hat{\sigma}(\overline{X}|I^*)$ would be 0.0138(1.74), 0.0084(1.05), and 0.0130(1.64), respectively. As predicted by the equations in Table 2, increasing the number of raters on which $\overline{X}$ is based <u>decreases</u> the expected variability of the distribution of mean scores.

All of the above results depend upon one assumption--namely, that whenever we generalize over a facet (raters or items), we assume that the facet encompasses an essentially infinite number of observations. Sometimes evaluators wish to generalize to a finite population of raters and/or a finite universe of items. In this case, the equations in Table 2 are no longer appropriate, and the Appendix provides two equivalent expressions for estimating the expected variance of $\overline{X}$ for a population of raters of any size, N_r, and a universe of items of any size, N_i.

The Nedelsky Procedure

As discussed in the introduction to this paper, there are both similarities and differences between the Nedelsky and Angoff procedures. The two procedures are similar in that, for each item and rater, they both provide a probability that a minimally competent examinee will get an item correct. The Angoff procedure directly elicits this probability from each rater, whereas the Nedelsky procedure involves inferring this probability from the number of distractors that a rater believes would be eliminated by a minimally competent examinee. Here we consider both aspects of the Nedelsky procedure, beginning with an analysis of the Nedelsky probabilities (data set 2), which parallels our previous analysis of the Angoff procedure. Then we examine the eliminated distractors for the Nedelsky procedure (data set 3).

Probabilities of Correct Responses

Table 4 reports some sample statistics for the Nedelsky procedure based upon data set 2, the Nedelsky probabilities of a correct response. We note that the mean cutting score, $\overline{X}$, is 0.5563 (70.09) for the Nedelsky procedure; whereas for the Angoff procedure, $\overline{X}$ is 0.6632 (83.56), as indicated in Table 1. Clearly, there is a substantial difference in mean scores for the two procedures. Furthermore, Tables 1 and 4 indicate that the standard deviation of the rater means for the Nedelsky procedure is approximately double the corresponding standard deviation for the Angoff procedure.

Insert Table 4 about here

Table 5 reports a generalizability analysis of the Nedelsky probabilities based upon the same model, assumptions, and sample sizes used in presenting the corresponding results for the Angoff procedure in Table 3. In comparing the Nedelsky results in Table 5 with the Angoff results in Table 3, we note that each of the random effects variance components [$\hat{\sigma}^2(\underline{r})$, $\hat{\sigma}^2(\underline{i})$, and $\hat{\sigma}^2(\underline{ri})$] for the Nedelsky procedure is considerably larger than the corresponding variance component for the Angoff procedure. This fact directly results in larger estimates of $\hat{\sigma}(\overline{X})$, $\hat{\sigma}(\overline{X}|\underline{R}^*)$, and $\hat{\sigma}(\overline{X}|\underline{I}^*)$, for the Nedelsky procedure. $\hat{\sigma}(\overline{X}|\underline{R}^*)$, for generalizing over items, is approximately the same for the two procedures. However, $\hat{\sigma}(\overline{X})$, for generalizing over both raters and items, is twice as large for the Nedelsky procedure. A similar statement holds for $\hat{\sigma}(\overline{X}|\underline{I}^*)$, when generalization is over raters only. In a later section, we examine these and other differences between the two procedures in more detail.

Insert Table 5 about here

Eliminated Distractors

One way of viewing the results presented thus far is that, in terms of setting a single cutting score with the Nedelsky or Angoff procedure, $\hat{\sigma}^2(\underline{ri})$ is always a source of error, $\hat{\sigma}^2(\underline{r})$ is a source of error if generalization is over raters, and $\hat{\sigma}^2(\underline{i})$ is a source of error if generalization is over items. This statement is based upon the linear model in Equation 1 for the probability

assigned by a rater to an item. In the Nedelsky procedure, however, the data that are actually collected are eliminated distractors, not probabilities; even though the cutting score resulting from the Nedelsky procedure is based directly upon probabilities. (Technically, the cutting score is a linear function of the inferred probabilities, and a nonlinear function of the eliminated distractors.)

Several interesting, but potentially confounding, issues arise when we consider the set of eliminated distractors for raters and items. One of these issues is discussed below, and other issues are treated later. For a given item, if two raters indicate that the same number of distractors could be eliminated, then the (inferred) probability for these two raters will be the same, whether or not the raters agree on which distractors could be eliminated. Technically, in terms of the way Nedelsky formulated his procedure, such disagreement among raters has no bearing upon the cutting score that results from the procedure. However, it seems reasonable to believe that one's confidence in the Nedelsky procedure, in a specific context, might be influenced by the extent to which raters agree not only with respect to the number of distractors eliminated, but also with respect to which distractors could be eliminated.

To examine this issue, variance components can be estimated for a design in which raters are crossed with items, and distractors, d, are nested within items. We denote this design r x (d:i).[1] Formulas for estimating variance components for this design are presented in Table 6, along with the estimated variance components for data set 3. It is usual in many applications of generalizability theory to report random effects variance components, based

on the assumption that the population (or universe) size for each facet is essentially infinite. In this case, however, it seems unreasonable to consider the $n_d = 3$ distractors associated with each item as a sample from an essentially infinite universe of possible distractors for the item. Therefore, in Table 6, the variance components are reported under the assumption that distractors are fixed, and this assumption is indicated by the notation D^* .

Insert Table 6 about here

Let us concentrate on the two variance components in Table 6 that involve variability attributable to distractors. The variance component $\hat{\sigma}^2(d{:}i|D^*)$ reflects the average, over items, of the variance attributable to the proportion of raters who eliminate each distractor. The magnitude of $\hat{\sigma}^2(d{:}i|D^*)$ will be large when, on the average, raters judge an item's distractors to vary in their difficulty, or attractiveness, to examinees. By contrast, the magnitude of $\hat{\sigma}^2(rd{:}i|D^*)$ reflects disagreement or variability among raters in their judgments of distractor attractiveness for an item. To put it another way, the magnitude of $\hat{\sigma}^2(rd{:}i|D^*)$ reflects the extent to which raters disagree in their judgments about which distractors could be eliminated.

If we consider $\hat{\sigma}^2(d{:}i|D^*) = 0.0629$ as an estimate of "true" variability among distractors, then our estimate of "error" for $n_r = 5$ raters is:

$$\hat{\sigma}^2(rd{:}i\ D^*)/n_r = 0.1814/5 = 0.0363.$$

Evidently, the "error" variance (attributable to the differential attractiveness of distractors for different raters) is almost fifty percent as large as the

"true" variance among distractors. This suggests that, for these data, even when raters agree on the number of distractors that can be eliminated, there are substantial differences among raters concerning which distractors can be eliminated.

Eliminating Correct Alternative

In conducting a study with the Nedelsky procedure, it is usual to provide complete items, including the correct alternatives, to each rater. If the correct alternatives for all items are specified for the raters, then it is reasonable to expect that no rater would eliminate the correct alternative for any item--assuming, of course, that the items are well-constructed and the raters take their task seriously.

On the other hand, if the correct answer is not specified for the raters, then perhaps some raters will eliminate the correct alternative for some items. This is indeed what happened in this study. Specifically, the numbers of correct answers eliminated by raters 1 to 5 were 11, 9, 26, 14, and 16, respectively. We found no evidence of clerical error, or mis-keyed items to explain these results, and we have no reason to question the extent to which raters took their task seriously. However, it is likely that individual raters had differing degrees of familiarity with the content tested by specific items.

When a rater indicates that the correct answer could be eliminated by a minimally competent examinee, one could argue that the (inferred) probability assigned by the rater to the item should be zero, no matter how many distractors are eliminated by the rater. However, for the purposes of this study, we did not adhere to this argument. Rather, we followed Nedelsky's procedure, as he described it, and assigned probabilities on the basis of eliminated distractors only. It is interesting to note that if we had assigned a probability of zero whenever a rater eliminated the correct alternative, $\bar{X}$ would decrease and estimates of variability would increase.

A Comparison of the Two Procedures

Since the Nedelsky and Angoff procedures were both applied to the same items by the same raters, the data from these two procedures (data sets 1 and 2) can be analyzed jointly in a single design. Specifically, the appropriate analysis involves the $\underline{p} \times \underline{r} \times \underline{i}$ design, in which the two procedures ($\underline{p}$) are crossed with both raters and items. Table 7 provides equations for estimating the variance components for this design, and Table 8 provides the numerical values of these estimated variance components for our data.

Insert Tables 7 and 8 about here

The variance components, identified as $\hat{\sigma}^2(\alpha)$ in Table 8 are obtained by letting $N_{\underline{p}}$ approach infinity in Table 7; and these are called random effects variance components. The variance components identified as $\hat{\sigma}^2(\alpha|\underline{P}^*)$ in Table 8 are obtained by letting $n_{\underline{p}} = N_{\underline{p}}$ in Table 7; and these variance components are based on the assumption that procedures are fixed. The variance components $\hat{\sigma}^2(\alpha|\underline{P}^*)$ are appropriate when we restrict our interest to the actual procedures in our study. Strictly speaking, here, the variance components $\hat{\sigma}^2(\alpha|\underline{P}^*)$ seem more appropriate than the random effects variance components, $\hat{\sigma}^2(\alpha)$, because it seems difficult to consider these two procedures as a sample from some very large set of similar cutting score procedures. However, the random effects variance components are very useful in illustrating relationships between results for the $\underline{p} \times \underline{r} \times \underline{i}$ design and the two $\underline{r} \times \underline{i}$ designs discussed previously.

Tables 7 and 8 also provide equations and numerical values for estimates of the variability of $\bar{X}$, where $\bar{X}$ is, in this case, the mean over raters, items, and procedures. For example, Table 8 reports that $\bar{X}$ (over procedures) is .6097 (78.82), which is the mean of the $\bar{X}$'s reported in Tables 1 and 4.

The reader should note, however, that the estimates of the variability of $\bar{X}$ in Table 8 are not averages of the corresponding estimates in Tables 3 and 5. For example, $\hat{\sigma}(\bar{X}|P^*) = 0.0195$ (2.46), which is similar to $\hat{\sigma}(\bar{X}) = 0.0182$ (2.29) in Table 3 for the Angoff procedure, but quite different from $\hat{\sigma}(\bar{X}) = 0.0336$ (4.24) in Table 5 for the Nedelsky procedure. This pattern of results also holds for $\hat{\sigma}(\bar{X}|P^*,R^*)$ and $\hat{\sigma}(\bar{X}|P^*,I^*)$. One inference that might be drawn from these observations is that there would be no particular advantage in actually setting a cutting score by averaging $\bar{X}$ from both procedures--assuming one is primarily interested in minimizing the variability of $\bar{X}$.

Perhaps the most interesting result in Table 8 is that the variance components that contain p are relatively large, indicating that there are substantial differences between the two procedures and the probabilities that result from them. For example, $\hat{\sigma}^2(p|P^*)$ is about four times larger than $\hat{\sigma}^2(r|P^*)$, suggesting that there is considerably more variability attributable to differences in procedure means than to differences in rater means (over procedures). From another perspective, it can be shown that the observed variance in the two procedure means is:

$$\hat{\sigma}^2(\bar{X}_p) = \hat{\sigma}^2(p) + \frac{\hat{\sigma}^2(pr)}{n_r} + \frac{\hat{\sigma}^2(pi)}{n_i} + \frac{\hat{\sigma}^2(pri)}{n_i n_i} .$$

In other words, the variance components that contain p contribute directly to the disparity we have identified in the procedure means.

The results reported in Table 8 are based upon the same data that led to the results in Tables 3 and 5, for the two procedures separately. It might seem, therefore, that there ought to be some relationships between the variance components in Table 8 and those in Tables 3 and 5. This is indeed the case. The reader can verify that the average of the variance components for raters in Tables 3 and 5 is:

$$[\hat{\sigma}^2_1(\underline{r}) + \hat{\sigma}^2_2(\underline{r})]/2 = \hat{\sigma}^2(\underline{r}) + \hat{\sigma}^2(\underline{pr});$$

where variance components to the right of the equality are for the $\underline{p} \times \underline{r} \times \underline{i}$ design in Table 8, $\hat{\sigma}^2_1(r)$ is the variance component for raters, for the Angoff procedure in Table 3, and $\hat{\sigma}^2_2(r)$ is the variance component for raters, for the Nedelsky procedure, in Table 5. Similarly,

$$[\hat{\sigma}^2_1(\underline{i}) + \hat{\sigma}^2_2(\underline{i})]/2 = \hat{\sigma}^2(\underline{i}) + \hat{\sigma}^2(\underline{pi}), \text{ and}$$

$$[\hat{\sigma}^2_1(\underline{ri}) + \hat{\sigma}^2_2(\underline{ri})]/2 = \hat{\sigma}^2(\underline{ri}) + \hat{\sigma}^2(\underline{pri}).$$

In effect, Table 8 crystallizes many of the differences between the two procedures evident in comparing Table 3 with Table 5.

Differences in Sample Statistics for Raters

We can also examine differences between the two procedures using the sample statistics reported in Tables 1 and 4. In examining these differences, we will occasionally point out (without proof) relationships between the results in Tables 1 and 4 and the generalizability analyses results in Tables 3, 5, and 8.

Correlations and Covariances Among Raters, Within Procedures. Using Tables 1 and 4, the reader can verify that the average of the rater intercorrelations for the Angoff procedure is 0.187, and the corresponding result for the Nedelsky procedure is 0.222. In terms of covariances, these averages are 0.0061 and 0.0125 for the Angoff and Nedelsky procedures, respectively. The magnitude of these average covariances is influenced by the degree to which similar probabilities are assigned to items. Indeed, the average of the rater covariances is simply $\hat{\sigma}^2_1(\underline{i})$ for the Angoff procedure, and $\hat{\sigma}^2_2(\underline{i})$ for the Nedelsky procedure. Evidently, there is more variability over items in the probabilities assigned using the Nedelsky procedure. We will see further evidence of this fact, below.

Rater Means. Figure 1 provides a scatterplot of the rater means (over items) for the Angoff procedure (see Table 1) and the Nedelsky procedure (see Table 4). The reader can verify that the correlation in Figure 1 is -0.052; and, it can be shown that the covariance (in terms of the random effects variance components in Table 8) is

$$\hat{\sigma}^2(\underline{r}) + \hat{\sigma}^2(\underline{ri})/n_{\underline{i}} = (-0.0002) + 0.0071/126 \doteq -0.0001.$$

Clearly, there is little, if any, linear relationship between the two procedures in terms of the five rater means.[2] Note that this result is not influenced by the difference in the grand means ($\overline{X}$'s) for the two procedures.

It appears from Figure 1, however, that there are two clusters of raters--Raters 2,3 and Raters 1, 4,5. Given the small numbers of raters involved, we hesitate to say that there is a strong correlation among raters within clusters;

however, Figure 1 certainly does not preclude this possibility. In any case, Raters 2 and 3 are outstanding in that they assign relatively low probabilities using the Nedelsky procedure and relatively high probabilities using the Angoff procedure.

Insert Figure 1 about here

Rater Standard Deviations. Figure 2 provides a scatterplot of the statistics $\hat{\sigma}_i(X_{ri})$ for each rater, by both procedures. Recall that, for a given rater and procedure, $\hat{\sigma}_i(X_{ri})$ is the standard deviation of the probabilities assigned to items. We observe that the standard deviations for the Nedelsky procedure are somewhat higher than those for the Angoff procedure, which is consistent with the fact that the variance components for items and interactions are higher for the Nedelsky procedure. Again, however, Rater 3 and, to some extent, Rater 2 appear to be different from the other three raters. Specifically, for both procedures, Raters 2 and 3 exhibit less variability in the probabilities they assign to items.

Insert Figure 2 about here

Differences in Sample Statistics for Items

In principal, we could construct tables, analogous to Tables 1 and 4, that would report, for both procedures, sample statistics for every item. Since we have 126 items, however, the resulting tables would be too large and too detailed to be very informative. Rather, we provide four perspectives on items statistics in two figures and two tables.

Figure 3 provides a frequency polygon for the average (over raters) of the probabilities assigned to items by both procedures; and Figure 4 provides a frequency polygon of the standard deviation of the probabilities assigned to items. Consistent with previously discussed results, Figure 3 indicates that the modal probability (interval) is considerably higher for the Angoff procedure. Also, consistent with previous results, Figure 4 indicates that there is somewhat more variability in the probabilities assigned to items using the Nedelsky procedure. Most importantly, however, the Nedelsky standard deviations in Figure 4 are bimodal. As discussed below, this bimodality is not an artifact of these data--it is a result that is virtually guaranteed by the Nedelsky procedure, per se.

Insert Figures 3 and 4 about here

Recall that, for each rater the probability assigned to an item by the Nedelsky procedure is the inverse of the number of non-eliminated alternatives. For the four-alternative items that characterize this study, this procedure for assigning probabilities means that the only (inferred) probabilities that can be assigned to an item using the Nedelsky procedure are 0.25, 0.33, 0.50, and 1.00. In particular, note that there can be no probability between 0.50 and 1.00. Now, consider the probabilities assigned by raters to an item. If all raters assign probabilities in the range 0.25 to 0.50, the standard deviation will be relatively small; and, of course, if they all assign probabilities of 1.00, the standard deviation will be zero. However, the standard deviation will be relatively large when some raters assign a probability of 1.00, and other raters assign probabilities of 0.50 or lower.

The bimodality in Figure 4, then, seems almost certainly a direct result of having only a small number of unequally spaced probabilities with the Nedelsky procedure. Furthermore, this peculiar characteristic of the probability scale is a plausible explanation for the fact that our estimates of the variability of $\bar{X}$ are higher for the Nedelsky procedure than for the Angoff procedure. (See Tables 3 and 5.) Also, the restricted nature of the Nedelsky probability scale may account for the differences in the means for the two procedures, at least to some extent. To examine these issues in more detail, let us consider Tables 9 and 10.

Insert Tables 9 and 10 about here

Tables 9 and 10 provide relative frequency distributions (over items) for the probabilities assigned using the Angoff and Nedelsky procedures, respectively. Inspection of these tables reveals several points of interest. First, no rater assigned probabilities below 0.20 using the Angoff procedure. This implies that the range of probabilities for the two procedures is about the same; and, consequently, differential restriction in range is not a factor of importance in this study. Second, for the Nedelsky procedure, on the average, probabilities below 0.50 were used for 28 percent of the items, whereas for the Angoff procedure, they were used for only 7 percent of the items. Third, for the Angoff procedure, on the average, probabilities in the range 0.60 to 0.95 were used with 53 percent of the items, whereas the Nedelsky procedure precluded use of such probabilities.

These points and visual inspection of Tables 9 and 10 reveal a consistent tendency for raters to assign more homogeneous probabilities using the Angoff procedure. Furthermore, it appears that a rater who uses a probability of 0.33 or 0.50 with the Nedelsky procedure is very likely to use a somewhat higher probability when given the opportunity to do so with the Angoff procedure.

Operationalizing Conceptions of Minimum Competence

There are many ways in which the Nedelsky and Angoff procedures appear to be very similar. For example, they both involve raters' judgments about individual items; they both yield, directly or indirectly, a matrix of rater-by-item probabilities, and, given this matrix, the computational process for

arriving at a cutting score is the same for both procedures. The procedures obviously differ in that probabilities are directly elicited in the Angoff procedure, whereas probabilities are inferred from eliminated distractors in the Nedelsky procedure.

It is also possible that the two procedures differ, to some extent, in the way they technically allow a rater to operationalize a conception of minimum competence. In the Angoff procedure, to arrive at a probability, a rater might conceptualize a group of minimally competent persons and reflect upon what proportion would get the item correct. Alternatively, for the Angoff procedure, a rater might conceptualize a single minimally competent person and reflect upon what proportion of the times this person would correctly respond to the item, if it were administered a large number of times.

For the Nedelsky procedure, however, there are only as many distinct probabilities that can be assigned (indirectly) as there are alternatives to the item, and these probabilities are not equally spaced. Logic suggests, therefore, that neither of the above two conceptualizations works very well with Nedelsky procedure. For example, if a rater believes that 75 percent of a group of minimally competent persons would get an item correct, the rater cannot eliminate some number of alternatives that will yield a probability of 0.75. Technically, the rater cannot even report the average number of alternatives that a group of minimally competent persons would eliminate, unless this number is an integer.

It seems, then, that the Nedelsky procedure constrains a rater to conceptualize minimum competency in terms of the performance of a single person on a single administration of an item, with the additional constraint that

this person will respond based upon a process of eliminating distractors. We know of no compelling empirical evidence to suggest that examinees (specifically, minimally competent examinees) generally respond to an item based upon a process of eliminating distractors, even though this process is frequently recommended to potential examinees. However, even if examinees do respond in this manner, there still seem to be relatively clear differences in the conceptualization of minimal competence implicit in the Angoff and Nedelsky procedures.

This study cannot directly address the extent to which different conceptualizations of minimum competency may have influenced the study's results; and it is not likely that raters gave this matter a great deal of conscious consideration. Nevertheless, any cutting score procedure necessitates some conceptualization of minimum competence; it seems likely that the conceptualizations are different for the Angoff and Nedelsky procedures; and evaluators are probably well-advised to consider such differences in choosing a cutting score procedure in a given context.

Cutting Scores other than $\bar{X}$

It is important to note that, throughout this paper, we have assumed that the cutting score, $\bar{X}$, that results from either procedure is the mean of $\bar{X}_r$, for all raters who participated in the study. For example, we pointed out that Raters 2 and 3, in our study, appear to be different from the other three raters. However, we did not suggest eliminating them from the study for the purposes of calculating a cutting score. In our opinion, unless there is clear evidence that a rater did not adhere to the intended procedure, it is probably not generally

advisable to eliminate atypical raters in determining the cutting score. (We assume, of course, that raters were chosen carefully in the first place.) However, if an atypical rater were eliminated, it would be best to redo analyses using the remaining raters, only. We make this suggestion because the elimination of an atypical rater, after the study is completed, probably implies that one has changed one's conceptualization of the intended population of raters.

Reconciliation Process. Sometimes, rather than using $\bar{X}$ as the cutting score, it is suggested that a cutting score be determined by a reconciliation process. In principal, such a process might be applied in conjunction with either the Nedelsky or the Angoff procedure. For example, after the five raters in this study completed the Angoff procedure, they were instructed, as a group, to reconcile their differences on each item. In Table 1 the mean (over items) of these reconciled ratings is denoted $\underline{r}(\underline{c})$. One typical result of using a reconciliation process is that certain raters tend to dominate, or unequally influence, the reconciled ratings. This is indeed what happened in our study, as indicated by the high correlations between the actual and reconciled ratings for Raters 1 and 2. The effect of this dominance by Raters 1 and 2 is that the reconciled cutting score (0.70) is quite a bit different from $\bar{X}$ (0.66).

There is certain logic in using a reconciliation process that appears to be compelling. One might argue that the ideal result of using either the Nedelsky or the Angoff procedure is for raters to agree on every item. Therefore, why not force them to concur? One argument against this logic is that forced consensus is not really agreement, although forced consensus may effectively hide disagreement. Also, we point out that a reconciliation process does not guarantee that the same cutting score will result each time a study is

replicated. If a study is replicated a large number of times, with different raters, the average reconciled cutting score might be considerably different from the average $\bar{X}$ or λ, in Equation 1; however, there could be as much, or even more, variability in the distribution of reconciled cutting scores as there is in the distribution of $\bar{X}$'s. We do not mean to imply, however, that a reconciliation procedure should be avoided, necessarily; rather, we wish to emphasize that use of reconciliation procedure involves complexities over and above those encompassed by either the Nedelsky or the Angoff procedure.

Nedelsky's Cutting Score. When Nedelsky originally described his procedure he did not actually suggest using $\bar{X}$ (or $n_i\bar{X}$) as a cutting score. Rather, the cutting score he suggested using is

$$M_{FD} + k\,\sigma_{FD} .$$

We find Nedelsky's discussion of M_{FD}, k, and σ_{FD} somewhat confusing. However, it appears that M_{FD} is intended to be the mean test score for a group of "border-line" examinees, only (Nedelsky, 1954, p. 5); σ_{FD} is the standard deviation of this distribution; and k is an a priori defined constant used to classify these "border-line" examinees into passing and failing examinees. Since Nedelsky suggests using our $n_i\bar{X}$ as an estimate of M_{FD}, it is clear that his cutting score will equal $n_i\bar{X}$ only if k is defined as zero of σ_{FD} is zero.

It is not clear to these authors why one would use $M_{FD} + \sigma_{FD}$ as a cutting score if one actually had test scores for a known group of "border-line" examinees. In such a case, the test data themselves would likely provide a reasonably

sound basis for defining a cutting score independent of raters' judgments. We infer, therefore, that Nedelsky probably wants us to consider a hypothetical group of borderline examinees. We have already argued that there may be a logical inconsistency in conceptualizing a group of minimally competent examinees when one uses the Nedelsky procedure. However, even if we overlook this issue, we are still faced with the problem of estimating σ^2_{FD} (a parameter for a test score distribution) using only the raters' probabilities.

It can be shown that the formula suggested by Nedelsky (1954, p. 12) for estimating σ^2_{FD} is

$$\hat{\sigma}^2_{FD} = \sum_r \sum_i X_{ri}(1 - X_{ri})/n_r$$

$$\doteq n_i[\bar{X}(1 - \bar{X}) - \hat{\sigma}^2(r) - \hat{\sigma}^2(i) - \hat{\sigma}^2(ri)];$$

where X_{ri} is the (inferred) probability assigned to item i by rater r. Nedelsky provides a rationale for his estimate of σ^2_{FD}; but, in our opinion, his rationale is weak in that it confounds considerations of parameters and estimates. However, even if one accepts his formula for estimating σ^2_{FD}, the very process of defining a cutting score as $M_{FD} + k\,\sigma_{FD}$ requires fairly strong assumptions and a substantial degree of subjective judgment over and above that required to estimate the cutting score $n_i\bar{X}$. Whether or not such complexity is advisable depends upon the specific context of the cutting score decision process; however, there are probably not many contexts in which this complexity is warranted and the procedure is easily defended.

Measurement Reliability or Dependability

In the Nedelsky and Angoff procedures the cutting score that results may be viewed as the mean, over items and raters, of the probabilities assigned to items. In this paper, we have denoted this mean as $\overline{X}$, and suggested that the magnitudes and sources of error in a cutting score procedure may be examined through studying the expected variance of $\overline{X}$. For both the Nedelsky and Angoff procedures, there are three possible estimates of this variance, depending upon whether a decision-maker wants to generalize to a population of raters and a universe of items, to a population of raters for a fixed set of items, or to a universe of items for a fixed set of raters.

More specifically, in generalizability theory the (observed mean) cutting score, $\overline{X}$, resulting from a particular study is an unbiased estimate of λ in Equation 1. (Recall that λ is the cutting score that would result if the population of raters used the Nedelsky or Angoff procedure with the universe of items.) However, we know that if a study were replicated a large number of times, it is very likely that the $\overline{X}$'s from these studies would vary; and any such variation reflects error in using $\overline{X}$ from a single study as an estimate of λ. It is usually not possible to conduct a number of replications of a cutting score procedure; but, even so, generalizability theory enables us to estimate the expected variance in the distribution of $\overline{X}$ (see Table 2). It is the expected variance of the distribution of means that we have examined in considerable detail. Again, however, there is not just one estimate of this variance. There are many estimates depending upon (a) whether one wishes to generalize over raters, items, or both; and (b) the sizes of the samples of raters and items used to calculate $\overline{X}$.

We emphasize that the numerical results reported in this paper are for a specific study, only; and, as such, these results are illustrative, rather than definitive. Nevertheless, there appear to be noticeable differences in the means (or cutting scores) for the two procedures. Also, for each procedure, there is evidence of error, as reflected in the expected variances of the distributions of means over replications; and these variances frequently have considerably different magnitudes for the two procedures. Given these results, it seems reasonable to consider their potential impact on issues of reliability, or measurement dependability. A complete discussion of these issues is beyond the intended scope of this paper. We will, however, consider these issues in the context of the index of dependability $\Phi(\lambda)$, defined by Equation 6 in Table 11.

This index was developed by Brennan and Kane (1977a) using generalizability theory and the linear model for the $\underline{p} \times \underline{i}$ design:

$$\underline{Y}_{\underline{pj}} = \mu + \mu_{\underline{p}}^{\sim} + \mu_{\underline{j}}^{\sim} + \mu_{\underline{pj}}^{\sim} . \qquad (7)$$

(See also Brennan, 1978, and Brennan and Kane, 1977b.) In Equation 7, $\underline{Y}_{\underline{pj}}$ is the observed score of person $\underline{p}$ on item $\underline{j}$, and the terms to the right of the equality are the score effects or components for the decomposition of $\underline{Y}_{\underline{pj}}$. The linear models in Equations 1 and 7 are formally identical. We have used different notation in each of them for the purpose of emphasizing that Equation 1 is applied to a rater-by-item matrix of probabilities, whereas Equation 2 is applied to the person-by-item matrix of observed scores.

For any meaningful joint use of Equations 1 and 7, the item universe must be the same for both equations although the effect for items in Equation 1, $\lambda_{\underline{i}}\sim$, is different from the effect for items in Equation 7, $\mu_{\underline{i}}\sim$. Most importantly, λ and μ in Equations 1 and 7 are very different. The parameter λ is the cutting score (or grand mean of the probabilities) for the population of raters and universe of items; whereas the parameter μ is the grand mean of the observed scores, $\underline{Y}_{pj}$, for the population of persons and the universe of items.

Using generalizability theory and the linear model in Equation 7, Brennan and Kane (1977a) derived Equation 8 in Table 11 as an estimate of their index of dependability. This estimate is identified as $\hat{\Phi}(\lambda)$ in Equation 8 to emphasize that it is based on the assumption that λ is somehow known, without error. This assumption is reflected in the term $(\overline{\underline{Y}} - \lambda)^2$ in the numerator and denominator of Equation 8. When λ is not known, however, and we use $\overline{\underline{X}}$ from a particular study as an estimate of λ, then this term is no longer appropriate.

Insert Table 11 about here

Furthermore, we may not simply replace λ with $\overline{\underline{X}}$ in the term $(\overline{\underline{Y}} - \lambda)^2$ because the expected value of a squared quantity is not equal to the square of the expected value. Rather, the expected value of $(\overline{\underline{Y}} - \overline{\underline{X}})^2$ is

$$\mathcal{E}_{\underline{R}}\mathcal{E}_{\underline{I}}(\overline{\underline{Y}} - \overline{\underline{X}})^2 = (\overline{\underline{Y}} - \lambda)^2 + \hat{\sigma}^2(\overline{\underline{X}}), \qquad (9)$$

if we wish to generalize over samples of raters ($\underline{R}$) and samples of items ($\underline{I}$). If we wish to generalize over samples of items, only, then

$$\mathcal{E}_{\underline{I}}(\overline{\underline{Y}} - \overline{\underline{X}})^2 = (\overline{\underline{Y}} - \lambda)^2 + \hat{\sigma}^2(\underline{X}|\underline{R}^*). \qquad (10)$$

Table 2 provides equations for $\hat{\sigma}^2(\overline{\underline{X}})$ and $\hat{\sigma}^2(\overline{\underline{X}}|\underline{R}^*)$ in terms of variance components for the effects in Equation 1; and Equations 9 and 10 can be derived using an approach employed by Brennan and Kane (1977a, p. 280).

It follows from Equations 9 and 10 that when we use $\overline{\underline{X}}$ as an estimate of λ, we should also subtract $\hat{\sigma}^2(\overline{\underline{X}})$ or $\hat{\sigma}^2(\overline{\underline{X}}|\underline{R}^*)$, as appropriate, from both the numerator and the denominator of Equation 8 in Table 11. The two resulting (modified) estimates of the index of dependability $\Phi(\lambda)$ are provided by Equations 11 and 12 in Table 11.

Let us return now to the original question that motivated our development of Equations 11 and 12--namely, what effect do different values for $\overline{\underline{X}}$ and its expected variability, for the Nedelsky and Angoff procedures, have on reliability or measurement dependability? Without loss of generality, we restrict ourselves to considering $\hat{\Phi}(\overline{\underline{X}})$ in Equation 11 for generalizing over raters and items. Since $\Phi(\lambda)$ can be no greater than one, decreasing the numerator and denominator in Equation 8 by $\hat{\sigma}^2(\overline{\underline{X}})$ results in decreasing the magnitude of the estimate of the Brennan-Kane index. This is to be expected, because we have introduced additional sources of error attributable to the procedure used to establish a cutting score.

Furthermore, "all other things" being equal, the larger the magnitude of $\hat{\sigma}^2(\overline{\underline{X}})$, the smaller the magnitude of $\hat{\Phi}(\overline{\underline{X}})$. Since our study results suggest that $\hat{\sigma}^2(\overline{\underline{X}})$ is larger for the Nedelsky procedure, we might expect $\hat{\Phi}(\overline{\underline{X}})$ to be smaller

for the Nedelsky procedure. However, "all other things" are not equal unless the cutting scores for the procedures are equal. When they are unequal, as we found, the magnitude of $(\overline{Y} - \overline{X})^2$ will be different; and this difference, in turn, will affect the magnitude of $\hat{\Phi}(\overline{X})$. Moreover, whether or not higher values of $\overline{X}$ will result in higher values of $(\overline{Y} - \overline{X})^2$ depends entirely upon the magnitude of $\overline{Y}$. In brief, the effect of $\overline{X}$ and $\hat{\sigma}^2(\overline{X})$ on the magnitude of $\hat{\Phi}(\lambda)$ cannot be predicted independent of test data for examinees; and, it is not necessarily the case that lower values of $\hat{\sigma}^2(\overline{X})$ are always associated with higher values for estimates of measurement dependability.

Note that we have not suggested that $\hat{\sigma}^2(\overline{X}|I^*)$ be considered in the context of modifying the Brennan-Kane index. Of course, there is an equation analogous to Equations 9 and 10--namely,

$$\mathcal{E}_R(\overline{Y} - \overline{X})^2 = (\overline{Y} - \lambda)^2 + \hat{\sigma}^2(\overline{X}|I^*),$$

in which generalization is over samples of raters only. However, in this equation, items are considered fixed; and if we incorporate $\hat{\sigma}^2(\overline{X}|I^*)$ into an estimate of $\Phi(\lambda)$ we must then consider items fixed in estimating the other variance components, too. To do so means that there is no larger universe of items (or tests) to which we wish to generalize; and, under such circumstances, estimates of reliability, generalizability, or dependability for the p x i design are usually undefined.

Summary and Conclusions

Based upon an application of generalizability theory to a rater-by-item matrix of probabilities, we have provided and discussed equations for estimating the expected variability in a cutting score determined by the Nedelsky or Angoff procedure. Our development assumes that the cutting score in a particular study is the observed mean (probability) over raters and items, and that this mean may be viewed as an estimate of an "idealized" cutting score, defined as the mean for a population of raters and a universe of items. In this sense, the expected variability of the observed mean is error variance attributable to a particular application of the procedure used to define a cutting score.

We have applied this approach to data sets resulting from the application of the Nedelsky and Angoff procedures by five raters to a 126-item test. Also, we have examined these results for each procedure separately, and we have compared results over procedures. Our data indicate that both the cutting scores and their expected variances are considerably different for the two procedures. We have postulated that these differences may be explained, in whole or in part, by differences in the ways probabilities are assigned using the two procedures, or differences in the ways minimum competency is conceptualized. Both explanations depend heavily on the fact that the Nedelsky procedure necessarily (although indirectly) restricts a rater to a small discrete number of unequally spaced probabilities.

In examining the two procedures, we have considered several issues not directly associated with variability in the mean cutting score, $\bar{X}$. Our data suggest, for example, that even when raters agree on the number of alternatives

to eliminate using the Nedelsky procedure, these same raters may disagree on which alternatives to eliminate. Also, we found that raters, using the Nedelsky procedure, eliminated a considerable number of correct alternatives. Furthermore, we have briefly discussed some issues associated with use of a reconciliation procedure and the elimination of atypical raters.

Finally, we have examined the influence of different values of $\bar{X}$, and the expected variance in the distribution of $\bar{X}$, on reliability, or measurement dependability. To do so, we developed a modification of the Brennan-Kane index of dependability, $\Phi(\lambda)$. We found that, for a given value of $\bar{X}$, increasing the expected variance of $\bar{X}$ results in decreasing the estimate of $\Phi(\lambda)$. However, if both $\bar{X}$ and its variance change, then the estimate of $\Phi(\lambda)$ could increase, decrease, or even remain unchanged--depending on results for examinee test data.

The numerical results reported in this paper are for a single study, only. As such, they surely do not form a sufficient basis for passing judgment on either the Nedelsky or the Angoff procedure. Even so, these data do suggest that the differences between these procedures may be of greater consequence than their apparent similarities. In particular, the restricted nature of the Nedelsky (inferred) probability scale may constitute a basis for rejecting this procedure in certain contexts.

Appendix

The r x i Design and Sampling from a Finite Universe

Table 2 provides equations for obtaining estimated random effects variance components for the r x i design. We emphasize that these variance components are for a random effects model in which both the size of the population of raters, N_r, and the size of the universe of items, N_i, are assumed to approach infinity, (i.e., $N_r \rightarrow \infty$ and $N_i \rightarrow \infty$). The equations in Table 2 for estimating the variability of mean scores do distinguish, however, between G study sample sizes and D study sample sizes. (D study sample sizes are identified with primes.)

The G study (i.e., generalizability study) sample sizes are the actual numbers of raters and items on which G study data are available; and these are the sample sizes used to calculate $\hat{\sigma}^2(r)$, $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(ri)$ in terms of mean squares. A decision-maker, however, may be interested in a D study involving consideration of the expected value of statistics, such as $\hat{\sigma}^2(\overline{X})$ for sample sizes that are different from the G study sample sizes. Of course, the G study and D study may be the same study, in which case there is no distinction between G study and D study sample sizes.

We wish to develop a general expression for the expected variance of the mean, $\overline{X}$, for samples of n'_r raters from a population of any size, N_r, and samples of n'_i items from a universe of any size, N_i. We begin by expressing expected mean squares in terms of variance components using the Cornfield and Tukey (1956) procedures (treated by Millman and Glass, 1967; Kirk, 1968; and others):

$$\mathcal{E}[\underline{MS}(\underline{r})] = (1 - \underline{n}_{\underline{i}}/\underline{N}_{\underline{i}})\sigma^2(\underline{ri}) + \underline{n}_{\underline{i}}\sigma^2(\underline{r}|\underline{N}_{\underline{i}}); \quad (A.1)$$

$$\mathcal{E}[\underline{MS}(\underline{i})] = (1 - \underline{n}_{\underline{r}}/\underline{N}_{\underline{r}})\sigma^2(\underline{ri}) + \underline{n}_{\underline{r}}\sigma^2(\underline{i}|\underline{N}_{\underline{r}}); \text{ and} \quad (A.2)$$

$$\mathcal{E}[\underline{MS}(\underline{ri})] = \sigma^2(\underline{ri}). \quad (A.3)$$

In Equations A.1 to A.3 it is important to note that $\sigma^2(\underline{r}|\underline{N}_{\underline{i}})$ is not identical to the random effects variance component $\sigma^2(\underline{r})$ unless $\underline{N}_{\underline{i}} \to \infty$; similarly, $\sigma^2(\underline{i}|\underline{N}_{\underline{r}})$ is not identical to $\sigma^2(\underline{i})$ unless $\underline{N}_{\underline{r}} \to \infty$; and $\sigma^2(\underline{ri})$ is unaffected by the size of $\underline{N}_{\underline{r}}$ and/or $\underline{N}_{\underline{i}}$. Also, note that the sample sizes in Equations A.1 to A.3 are for the G study--not the D study.

Now, in terms of estimates of the variance components in Equations A.1 to A.3,

$$\hat{\sigma}^2(\underline{\bar{X}}|\underline{N}_{\underline{r}},\underline{N}_{\underline{i}}) = \left(1 - \frac{\underline{n}'_{\underline{r}}}{\underline{N}_{\underline{r}}}\right)\frac{\hat{\sigma}^2(\underline{r}|\underline{N}_{\underline{i}})}{\underline{n}'_{\underline{r}}} + \left(1 - \frac{\underline{n}'_{\underline{i}}}{\underline{N}_{\underline{i}}}\right)\frac{\hat{\sigma}^2(\underline{i}|\underline{N}_{\underline{r}})}{\underline{n}'_{\underline{i}}}$$

$$+ \left(1 - \frac{\underline{n}'_{\underline{r}}}{\underline{N}_{\underline{r}}}\right)\left(1 - \frac{\underline{n}'_{\underline{i}}}{\underline{N}_{\underline{i}}}\right)\frac{\hat{\sigma}^2(\underline{ri})}{\underline{n}'_{\underline{r}}\underline{n}'_{\underline{i}}}. \quad (A.4)$$

Equation A.4 results from well-known principles concerning the variance of the distribution of means for (D study) samples of size $\underline{n}'$ randomly sampled from a popultion or universe of size $\underline{N}$ (see, for example, Cochran, 1977, p. 23). A slight modification of Equation A.4 is sometimes discussed in treatments of matrix sampling (see, for example, Sirotnik and Wellington, 1977, p. 354).

Equation A.4, however, is sometimes awkward to use for a particular D study, because G study results are usually reported in terms of estimated random effects variance components--not the estimated G study variance components $\hat{\sigma}^2(\underline{r}|\underline{N}_{\underline{i}})$ and $\hat{\sigma}^2(\underline{i}|\underline{N}_{\underline{r}})$, for sampling from a finite universe. Brennan (1977) has shown that

$$\hat{\sigma}^2(\underline{r}|\underline{N}_{\underline{i}}) = \hat{\sigma}^2(\underline{r}) + \hat{\sigma}^2(\underline{ri})/\underline{N}_{\underline{i}}, \tag{A.5}$$

and

$$\hat{\sigma}^2(\underline{i}|\underline{N}_{\underline{r}}) = \hat{\sigma}^2(\underline{i}) + \hat{\sigma}^2(\underline{ri})/\underline{N}_{\underline{r}}; \tag{A.6}$$

where estimated variance components to the right of the equalities are the random effects variance components in Table 2.

Given equations A.5 and A.6, Equation A.4 can be expressed as:

$$\hat{\sigma}^2(\overline{\underline{X}}|\underline{N}_{\underline{r}},\underline{N}_{\underline{i}}) = \left(1 - \frac{\underline{n}_{\underline{r}}'}{\underline{N}_{\underline{r}}}\right)\frac{\hat{\sigma}^2(\underline{r})}{\underline{n}_{\underline{r}}'} + \left(1 - \frac{\underline{n}_{\underline{i}}'}{\underline{N}_{\underline{i}}}\right)\frac{\hat{\sigma}^2(\underline{i})}{\underline{n}_{\underline{i}}'}$$

$$+ \left(1 - \frac{n_r' n_i'}{N_r N_i}\right)\frac{\hat{\sigma}^2(\underline{ri})}{n_r' n_i'} . \tag{A.7}$$

Equation A.7 can be used to obtain an unbiased estimate of the expected variance of $\overline{\underline{X}}$ for any values of $\underline{n}_{\underline{r}}'$, $\underline{n}_{\underline{i}}'$, $\underline{N}_{\underline{r}}$, and $\underline{N}_{\underline{i}}$ using random effects variance components, only. Let us consider, for example, the three special cases in Table 2. If $\underline{N}_{\underline{r}}$

and N_i both approach infinity, then Equation A.7 is $\hat{\sigma}^2(\bar{X})$ in Table 2. If N_r approaches infinity, and $n_i' = N_i$, then Equation A.7 is $\hat{\sigma}^2(\bar{X}|I^*)$ in Table 2. If N_i approaches infinity, and $n_r' = N_r$, then Equation A.7 is $\hat{\sigma}^2(\bar{X}|R^*)$ in Table 2.

References

Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.). Educational Measurement. Washington, D.C.: American Council on Education, 1971, 514-515.

Brennan, R. L. Generalizability analyses: Principles and procedures. ACT Technical Bulletin No. 26. Iowa City: The American College Testing Program, September, 1977.

Brennan, R. L. Some applications of generalizability theory to the dependability of domain-referenced tests. ACT Technical Bulletin No. 32. Iowa City: The American College Testing Program, April, 1979

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289. (a)

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. Psychometrika, 1977, 42, 609-625; Errata, 1978, 43, 289. (b)

Buck, L. A. Guide to the setting of appropriate cutting scores for written tests: A summary of the concerns and procedures. (Technical Memorandum 77-4, United States Civil Service Commission.) Washington, D. C.: U.S. Government Printing Office, 1977.

Cochran, W. G. Sampling techniques (3rd ed.). New York: Wiley, 1977.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, Calif.: Wadsworth, 1968.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. Review of Educational Research, 46, 1, 1976. 133-158.

Millman, J., & Glass, G. V. Rules of thumb for the ANOVA table. Journal of Educational Measurement, 1967, 4, 41-51.

National Council on Measurement in Education. Special issue on standard setting. Journal of Educational Measurement, 1978, 15, (4).

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Sirotnik, K. An analysis of variance framework for matrix sampling. Educational and Psychological Measurement, 1970, 30, 891-908.

Sirotnik, K., & Wellington, R. Incidence sampling: An integrated theory for "matrix sampling." Journal of Educational Measurement, 1977, 14, 343-399.

Zieky, M. J., & Livingston, S. A. Basic skills assessment, manual for setting standards. New Jersey: Educational Testing Service, 1977.

Footnotes

The junior author is currently Director, Southeastern Regional Test Development Center, Educational Testing Service, Atlanta, Georgia.

[1]For the $\underline{r} \times (\underline{d}:\underline{i})$ design, it can be argued that a rater may not examine a given distractor independent of the other distractors for an item. If so, then one independence assumption associated with the linear model for this design becomes suspect, at least to some extent. This issue, however, is relatively unimportant here because our analysis is intended only to summarize data that have an indirect bearing on the principal analyses in this paper.

[2]By definition, a variance component must be positive; however, estimates of variance components are occasionally negative. When a negative estimate occurs, sometimes it is advisable to treat it as zero (see Cronbach et al, 1972, and Brennan, 1977), and at other times it is best to leave the estimate unchanged (see Sirotnik, 1970). Here, we do not set $\hat{\sigma}^2(\underline{r})$ to zero because it is a mathematical fact that a covariance of the type in Figure 1 is exactly

$$\hat{\sigma}^2(\underline{r}) + \hat{\sigma}^2(\underline{ri})/n_{\underline{i}},$$

as shown by Cronbach et al. (1972, Chapter 8). This is true even if $\hat{\sigma}^2(\underline{r})$ is negative. Indeed, a negative covariance could never occur unless $\hat{\sigma}^2(\underline{r})$ and/or $\hat{\sigma}^2(\underline{ri})$ were negative.

Table 1

Means, Standard Deviations, and Intercorrelations

Among Raters for the Angoff Procedure

	Intercorrelations					Mean over items: $\bar{X}_{\underline{r}}$	S.D. over items: $\hat{\sigma}(X_{\underline{r}i})$
	$\underline{r}(2)$	$\underline{r}(3)$	$\underline{r}(4)$	$\underline{r}(5)$	$\underline{r}(c)$[a]		
r(1)	.525	.053	.046	.150	.731	0.6713	0.2033
r(2)		.171	.206	.382	.744	0.7194	0.1607
r(3)			.161	-.036	.237	0.6559	0.1193
r(4)				.209	.217	0.6167	0.1869
r(5)					.432	0.6525	0.2183
r(c)						0.6984	0.1541
	$\bar{X} = 0.6632$ (83.56)[b]					$\hat{\sigma}(\bar{X}_{\underline{r}}) = 0.0373$ (4.70)[b]	

[a] $\underline{r}(c)$ is a reconciled rating arrived at by the raters themselves.

[b] Numbers within parentheses are expressed in terms of number of items.

Table 2

Variance Components Notation and Formulas for

Random Effects $\underline{r}$ x $\underline{i}$ Design

Number of raters = $n_{\underline{r}}$ (G study); $n'_{\underline{r}}$ (D study)

Number of items = $n_{\underline{i}}$ (G study); $n'_{\underline{i}}$ (D study)

$$\hat{\sigma}^2(\underline{r}) = [MS(\underline{r}) - MS(\underline{ri})]/n_{\underline{i}}$$

$$\hat{\sigma}^2(\underline{i}) = [MS(\underline{i}) - MS(\underline{ri})]/n_{\underline{r}}$$

$$\hat{\sigma}^2(\underline{ri}) = MS(\underline{ri})$$

$$\hat{\sigma}^2(\bar{X}_{\underline{r}}) = \hat{\sigma}^2(\underline{r}) + \hat{\sigma}^2(\underline{ri})/n'_{\underline{i}}$$

$$\hat{\sigma}^2(\bar{X}) = \hat{\sigma}^2(\underline{r})/n'_{\underline{r}} + \hat{\sigma}^2(\underline{i})/n'_{\underline{i}} + \hat{\sigma}^2(\underline{ri})/n'_{\underline{r}}n'_{\underline{i}} \quad (3)$$

$$\hat{\sigma}^2(\bar{X}|R^*) = \hat{\sigma}^2(\underline{i})/n'_{\underline{i}} + \hat{\sigma}^2(\underline{ri})/n'_{\underline{r}}n'_{\underline{i}} \quad (4)$$

$$\hat{\sigma}^2(\bar{X}|I^*) = \hat{\sigma}^2(\underline{r})/n'_{\underline{r}} + \hat{\sigma}^2(\underline{ri})/n'_{\underline{r}}n'_{\underline{i}} \quad (5)$$

Table 3

ANOVA, Variance Components, and the Variability of Mean Scores for the Angoff Procedure

Effect (α)	df	SS	MS	$\hat{\sigma}^2(\alpha)$
r	4	0.7000	0.1750	0.0012
i	125	7.1443	0.0572	0.0061
ri	500	13.3526	0.0267	0.0267

$\hat{\sigma}^2(\bar{X}_r)$ = 0.0014		$\hat{\sigma}(\bar{X}_r)$ = 0.0373	(4.70)
$\hat{\sigma}^2(\bar{X})$ = 0.00033		$\hat{\sigma}(\bar{X})$ = 0.0182	(2.29)
$\hat{\sigma}^2(\bar{X}\|R^*)$ = 0.00009		$\hat{\sigma}(\bar{X}\|R^*)$ = 0.0095	(1.20)
$\hat{\sigma}^2(\bar{X}\|I^*)$ = 0.00028		$\hat{\sigma}(\bar{X}\|I^*)$ = 0.0167	(2.10)

Note. The terms $\hat{\sigma}^2(\alpha)$ are, more specifically, $\hat{\sigma}^2(r)$, $\hat{\sigma}^2(i)$, $\hat{\sigma}^2(ri)$. Results in the second half of this table for the variability of mean scores assume that $n'_r = n_r = 5$ and $n'_i = n_i = 126$.

Results within parentheses are expressed in terms of number of items.

Table 4

Means, Standard Deviations, and Intercorrelations Among Raters for Probability of Correct Response from Nedelsky Procedure

	Intercorrelations				Mean over items: $\bar{X}_{\underline{r}}$	S.D. over items: $\hat{\sigma}(X_{\underline{r}i})$
	$\underline{r}(1)$	$\underline{r}(3)$	$\underline{r}(4)$	$\underline{r}(5)$		
$\underline{r}(1)$	.307	.118	.196	.377	0.6438	0.2828
$\underline{r}(2)$		.065	.204	.350	0.5337	0.2317
$\underline{r}(3)$			.161	.195	0.4495	0.1827
$\underline{r}(4)$				.242	0.5700	0.2376
$\underline{r}(5)$					0.5844	0.2310
	$\bar{X} = 0.5563$ (70.09)[a]				$\hat{\sigma}(\bar{X}_{\underline{r}}) = 0.0717$ (9.03)[a]	

[a] Results within parentheses are expressed in terms of number of items.

Table 5

ANOVA, Variance Components, and Variability of Mean Scores for Probability of Correct Response from Nedelsky Procedure

Effect (α)	df	SS	MS	$\hat{\sigma}^2(\alpha)$
r	4	2.5890	0.6473	0.0048
i	125	13.1817	0.1055	0.0125
ri	500	21.4284	0.0429	0.0429

$\hat{\sigma}^2(\bar{X}_r) = 0.0051$	$\hat{\sigma}(\bar{X}_r) = 0.0717$	(9.03)
$\hat{\sigma}^2(\bar{X}) = 0.0011$	$\hat{\sigma}(\bar{X}) = 0.0336$	(4.24)
$\hat{\sigma}^2(\bar{X}\|R^*) = 0.0002$	$\hat{\sigma}(\bar{X}\|R^*) = 0.0130$	(1.64)
$\hat{\sigma}^2(\bar{X}\|I^*) = 0.0010$	$\hat{\sigma}(\bar{X}\|I^*) = 0.0321$	(4.04)

Note. The terms $\hat{\sigma}^2(\alpha)$ are more specifically $\hat{\sigma}^2(r)$, $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(ri)$. Results in the second half of this table, for the variability of mean scores, assume that $n'_r = n_r = 5$ and $n'_i = n_i = 126$.

Table 6

ANOVA and Variance Components for Eliminated Distractors Using Nedelsky Procedure

Effect (α)	df	SS	MS	$\hat{\sigma}^2(\alpha \mid D^*)$
r	4	9.3418	2.3354	0.0058
i	125	42.6245	0.3410	0.0121
d:i	252	124.9251	0.4957	0.0629
ri	500	79.9844	0.1600	0.0533
rd:i	1008	182.8853	0.1814	0.1814

$\hat{\sigma}^2(r \mid D^*) = [MS(r) - MS(ri)]/n_i n_d$

$\hat{\sigma}^2(i \mid D^*) = [MS(i) - MS(ri)]/n_r n_d$

$\hat{\sigma}^2(d:i \mid D^*) = [MS(d:i) - MS(rd:i)]/n_r$

$\hat{\sigma}^2(ri \mid D^*) = MS(ri)/n_d$

$\hat{\sigma}^2(rd:i \mid D^*) = MS(rd:i)$

Mean over items of proportion of distractors eliminated for raters 1 to 5:

$\bar{X}_r$ = 0.6984, 0.6085, 0.5053, 0.6561, 0.6878

$\bar{X}$ = 0.6312 $\qquad \hat{\sigma}(\bar{X}_r)$ = 0.0786

Table 7

Equations for Estimating Variance Components and the Expected Variance of the Mean Score for the $\underline{p} \times \underline{r} \times \underline{i}$ Design

Effect	Estimated Variance Component
$\underline{p}$	$[\underline{MS}(p) - \underline{MS}(pr) - \underline{MS}(pi) + \underline{MS}(pri)]/n_{\underline{r}} n_{\underline{i}}$
$\underline{r}$	$\{\underline{MS}(r) - \underline{MS}(ri) - (1 - n_{\underline{p}}/N_{\underline{p}})[\underline{MS}(pr) - \underline{MS}(pri)]\}/n_{\underline{p}} n_{\underline{i}}$
$\underline{i}$	$\{\underline{MS}(i) - \underline{MS}(ri) - (1 - n_{\underline{p}}/N_{\underline{p}})[\underline{MS}(pi) - \underline{MS}(pri)]\}/n_{\underline{p}} n_{\underline{r}}$
$\underline{pr}$	$[\underline{MS}(pr) - \underline{MS}(pri)]/n_{\underline{i}}$
$\underline{pi}$	$[\underline{MS}(pi) - \underline{MS}(pri)]/n_{\underline{i}}$
$\underline{ri}$	$[\underline{MS}(ri) - (1 - n_{\underline{p}}/N_{\underline{p}})\underline{MS}(pri)]/n_{\underline{p}}$
$\underline{pri}$	$\underline{MS}(pri)$

When $n_{\underline{p}} = N_{\underline{p}}$, we identify the variance components as $\hat{\sigma}^2(\alpha|\underline{P}^*)$. In terms of these variance components:

$$\hat{\sigma}^2(\bar{X}_{\underline{r}}|\underline{P}^*) = \hat{\sigma}^2(\underline{r}|\underline{P}^*) + \hat{\sigma}^2(\underline{ri}|\underline{P}^*)/n'_{\underline{i}}$$

$$\hat{\sigma}^2(\bar{X}|\underline{P}^*) = \hat{\sigma}^2(\underline{r}|\underline{P}^*)/n'_{\underline{r}} + \hat{\sigma}^2(\underline{i}|\underline{P}^*)/n'_{\underline{i}} + \hat{\sigma}^2(\underline{ri}|\underline{P}^*)/n'_{\underline{r}} n'_{\underline{i}}$$

$$\hat{\sigma}^2(\bar{X}|\underline{P}^*,\underline{R}^*) = \hat{\sigma}^2(\underline{i}|\underline{P}^*)/n'_{\underline{i}} + \hat{\sigma}^2(\underline{ri}|\underline{P}^*)/n'_{\underline{r}} n'_{\underline{i}}$$

$$\hat{\sigma}^2(\bar{X}|\underline{P}^*,\underline{I}^*) = \hat{\sigma}^2(\underline{r}|\underline{P}^*)/n'_{\underline{r}} + \hat{\sigma}^2(\underline{ri}|\underline{P}^*)/n'_{\underline{r}} n'_{\underline{i}}$$

Table 8

ANOVA and Variance Components for Probability of Correct Response with Both Procedures

Effect (α)	df	SS	MS	$\hat{\sigma}^2(\alpha)$	$\hat{\sigma}^2(\alpha \mid \underline{P}^*)$
p	1	3.5989	3.5989	0.0050	0.0050
r	4	1.5537	0.3884	-0.0002	0.0014
i	125	15.6601	0.1253	0.0074	0.0083
pr	4	1.7354	0.4339	0.0032	0.0032
pi	125	4.6652	0.0373	0.0019	0.0019
ri	500	20.9520	0.0419	0.0071	0.0210
pri	500	13.8285	0.0277	0.0277	0.0277

Means over procedures and items for raters 1 to 5:

$\overline{X}_{\underline{r}}$ = 0.6576, 0.6266, 0.5527, 0.5934, 0.6185

$\overline{X}$ = 0.6097 (78.82) $\quad \hat{\sigma}(\overline{X}_{\underline{r}}) = \hat{\sigma}(\overline{X}_{\underline{r}} \mid \underline{P}^*) = 0.0396$ (4.99)

$\hat{\sigma}^2(\overline{X} \mid \underline{P}^*)$	= 0.00038	$\hat{\sigma}(\overline{X} \mid \underline{P}^*)$	= 0.0195 (2.46)
$\hat{\sigma}^2(\overline{X} \mid \underline{P}^*, \underline{R}^*)$	= 0.00010	$\hat{\sigma}(\overline{X} \mid \underline{P}^*, \underline{R}^*)$	= 0.0100 (1.26)
$\hat{\sigma}^2(\overline{X} \mid \underline{P}^*, \underline{I}^*)$	= 0.00031	$\hat{\sigma}(\overline{X} \mid \underline{P}^*, \underline{I}^*)$	= 0.0176 (2.22)

Table 9

Relative Frequency Distribution by Raters by Probability of Correct Response Using Angoff Procedure

Probability of Correct Response[a]	Relative Frequency					
	$\underline{r}(1)$	$\underline{r}(2)$	$\underline{r}(3)$	$\underline{r}(4)$	$\underline{r}(5)$	Average
<0.20	0.00	0.00	0.00	0.00	0.00	0.00
(0.20, 0.25)	0.03	0.02	0.00	0.07	0.06	0.04
(0.30, 0.35)	0.02	0.00	0.00	0.02	0.06	0.02
(0.40, 0.45)	0.00	0.00	0.06	0.00	0.00	0.01
(0.50, 0.55)	0.42	0.24	0.19	0.44	0.36	0.33
(0.60, 0.65)	0.01	0.04	0.31	0.00	0.01	0.07
(0.70, 0.75)	0.26	0.33	0.23	0.37	0.31	0.30
(0.80, 0.85)	0.01	0.23	0.20	0.00	0.02	0.09
(0.90, 0.95)	0.12	0.07	0.01	0.10	0.04	0.07
>0.95	0.13	0.08	0.00	0.00	0.15	0.07
$\bar{X}_{\underline{r}}$	0.67	0.72	0.66	0.62	0.65	0.66

[a] Raters were constrained to report their probabilities in units of 0.05.

Table 10

Relative Frequency Distribution by Raters for Probability of Correct Response Using Nedelsky Procedure

Probability of Correct Response	Relative Frequency					Average
	r(1)	r(2)	r(3)	r(4)	r(5)	
0.25	0.06	0.06	0.07	0.05	0.00	0.05
0.33	0.16	0.25	0.42	0.16	0.17	0.23
0.50	0.40	0.52	0.43	0.57	0.60	0.50
1.00[a]	0.38	0.18	0.08	0.22	0.23	0.22
$\bar{X}_{\bar{r}}$	0.64	0.53	0.45	0.57	0.58	0.56

[a]Our analyses of these data used a probability of 0.99, rather than 1.00, for coding convenience.

Table 11

Equations for the Brennan-Kane Index of Dependability

and a Modification

$$\Phi(\lambda) = \frac{\sigma^2(\underline{p}) + (\mu - \lambda)^2}{\sigma^2(\underline{p}) + (\mu - \lambda)^2 + \sigma^2(\Delta)} \tag{6}$$

$$\hat{\Phi}(\lambda) = \frac{\hat{\sigma}^2(\underline{p}) + (\overline{\underline{Y}} - \lambda)^2 - \hat{\sigma}^2(\overline{\underline{Y}})}{\hat{\sigma}^2(\underline{p}) + (\overline{\underline{Y}} - \lambda)^2 - \hat{\sigma}^2(\overline{\underline{Y}}) + \hat{\sigma}^2(\Delta)} \tag{8}$$

where

$$\hat{\sigma}^2(\underline{p}) = [\underline{MS}(\underline{p}) - \underline{MS}(\underline{pj})]/\underline{n}_j$$

$$\hat{\sigma}^2(\underline{j}) = [\underline{MS}(\underline{j}) - \underline{MS}(\underline{pj})]/\underline{n}_p$$

$$\hat{\sigma}^2(\underline{pj}) = \underline{MS}(\underline{pj})$$

$$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(\underline{j})/\underline{n}'_j + \hat{\sigma}^2(\underline{pj})/\underline{n}'_j$$

$$\hat{\sigma}^2(\overline{\underline{Y}}) = \hat{\sigma}^2(\underline{p})/\underline{n}'_p + \hat{\sigma}^2(\underline{j})/\underline{n}'_j + \hat{\sigma}^2(\underline{pj})/\underline{n}'_p\underline{n}'_j$$

$$\hat{\Phi}(\overline{\underline{X}}) = \frac{\hat{\sigma}^2(\underline{p}) + (\overline{\underline{Y}} - \overline{\underline{X}})^2 - \hat{\sigma}^2(\overline{\underline{Y}}) - \hat{\sigma}^2(\overline{\underline{X}})}{\hat{\sigma}^2(\underline{p}) + (\overline{\underline{Y}} - \overline{\underline{X}})^2 - \hat{\sigma}^2(\overline{\underline{Y}}) - \hat{\sigma}^2(\overline{\underline{X}}) + \hat{\sigma}^2(\Delta)} \tag{11}$$

$$\hat{\Phi}(\overline{\underline{X}}|\underline{R}^*) = \frac{\hat{\sigma}^2(\underline{p}) + (\overline{\underline{Y}} - \overline{\underline{X}})^2 - \hat{\sigma}^2(\overline{\underline{Y}}) - \hat{\sigma}^2(\overline{\underline{X}}|\underline{R}^*)}{\hat{\sigma}^2(\underline{p}) + (\overline{\underline{Y}} - \overline{\underline{X}})^2 - \hat{\sigma}^2(\overline{\underline{Y}}) - \hat{\sigma}^2(\overline{\underline{X}}|\underline{R}^*) + \hat{\sigma}^2(\Delta)} \tag{12}$$

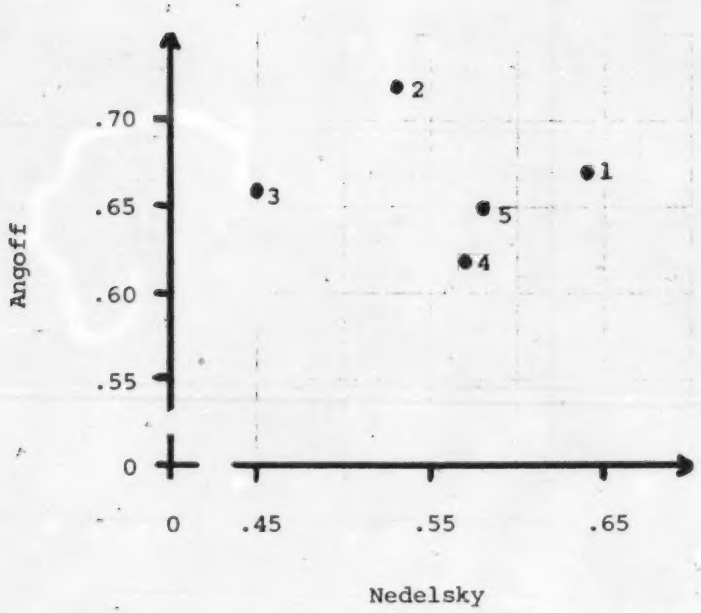

Figure 1. Rater means (over items) for probability of a correct response using Nedelsky and Angoff procedures.

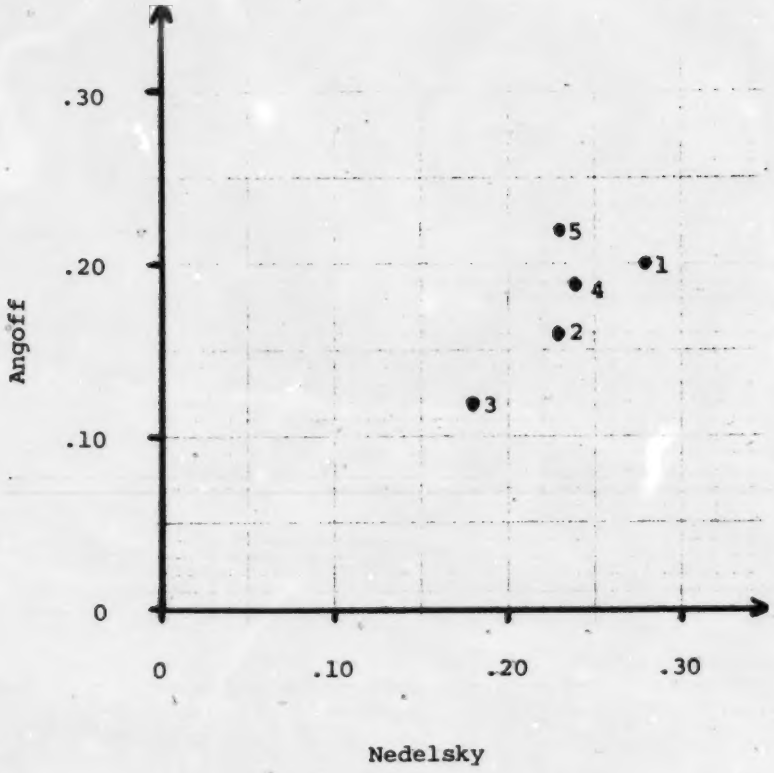

Figure 2. Standard deviations, for each rater, of the probabilities assigned to items using Nedelsky and Angoff procedures.

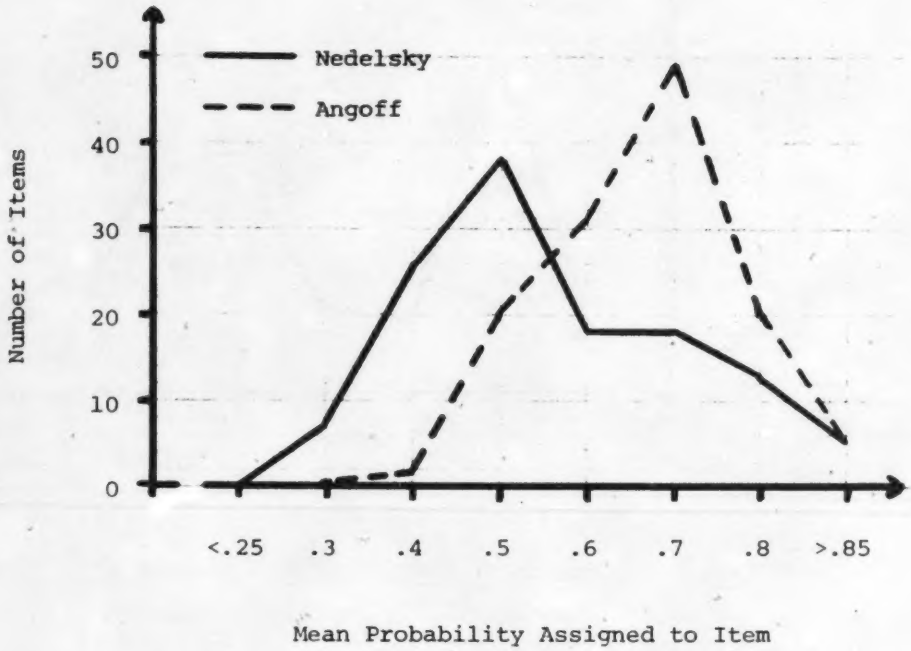

Figure 3. Frequency polygon of the means (over raters) of the probabilities assigned to items. Frequencies are plotted for the midpoints of intervals having width 0.10.

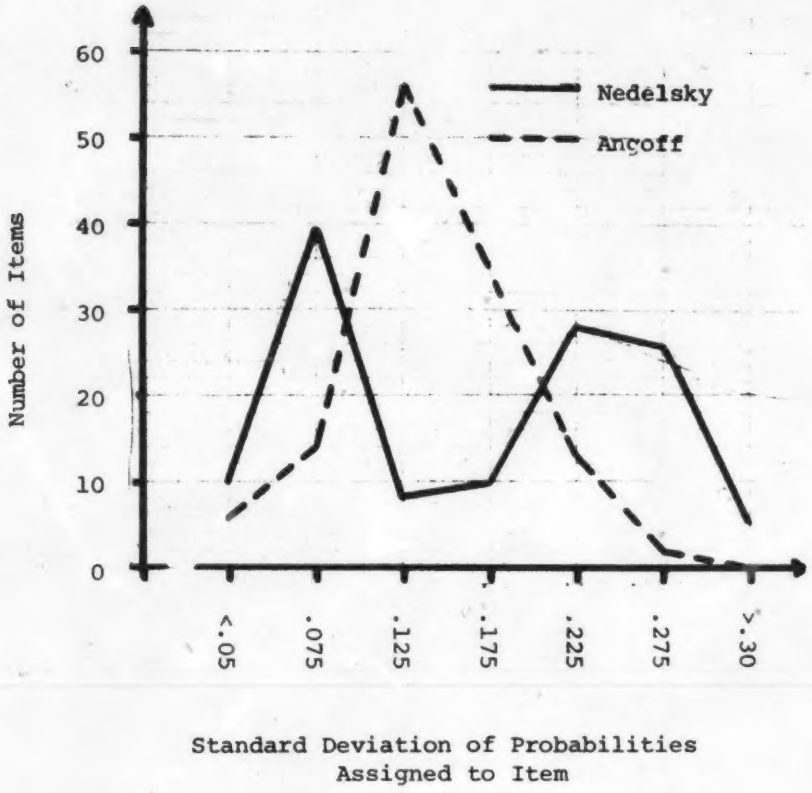

Figure 4. Frequency polygon of the standard deviations (over raters) of the probabilities assigned to items. Frequencies are plotted for the midpoints of intervals having width 0.05.